공학석사학위논문

다변량 시계열 데이터의 시각화를 위한 변수 인덱싱

Feature Indexing for Visualizing Multivariate Time Series Data

2019 년 7 월

서울대학교 대학원

산업공학과

박 봉 준

다변량 시계열 데이터의 시각화를 위한 변수 인덱싱

Feature Indexing for Visualizing Multivariate Time Series Data

지도교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함

2019 년 7 월

서울대학교 대학원

산업공학과

박 봉 준

박봉준의 공학석사 학위논문을 인준함

2019 년 7 월

위 원 장 홍 성 필 (인)

부위원장 조 성 준 (인)

위 원 박 종 헌 (인)

초록

제조 공정에서는 수많은 센서로부터 다양한 시계열 데이터가 발생한다. 이렇게 얻는 다변량 시계열 데이터로부터 공정의 성능과 이상을 예측한다. 예측된 공정의 이상은 원인 분석을 통해 적절한 조치로 이어진다. 본 연구는 다변량 시계열 데이터로부터 데이터 분석가가 쉽게 원인 분석을 할 수 있도록 변수 인덱싱을 통한 시각화 전처리하는 방법을 제안한다. 다변량 시계열 데이터를 각각 0-1 스케일링한 뒤, 이들 간의 상관계수를 링크의 비용으로 하는 완전 그래프로 변환하였다. 완전 그래프의 모든 노드를 지나는 최소 비용 경로를 정수 최적화 수식화하여 구하고 최적해를 기반으로 재배열, 반전을 통해 변수를 인덱싱하였다. 제안한 방법을 제철 공정 데이터에 적용하여 직사각형태의 단 채널 이미지로 변환한 뒤 무작위로 노이즈를 추가하여 이를 합성곱 신경망과 데이터 분석가를 통해 찾는 실험을 하였다. 전처리 전과 비교하여 합성곱 신경망의 분류 성능과 학습 속도가 향상되었고, 데이터 분석가가 노이즈의 위치를 더욱 정확히 찾을 수 있었다. 본 연구에서 제안한 전처리 방법이 합성곱 신경망의 성능과 인간의 데이터 분석 능력을 향상 시킴을 확인하였다.

주요어: 제조 공정, 다변량 시계열 데이터, 정수최적화, 합성곱 신경망, 전처리

학번: 2017-24488

목차

표 목차

그림 목차

제 1 장 서론

제조 공정에서는 진행되는 상태를 모니터링하기 위하여 온도, 압력 등을 측정하는 센서를 통해 데이터를 수집한다. 공정에 따라 수십에서 수백 개의 센서를 달고 시간별로 측정하여 데이터를 수집한다. 이렇게 발생하는 다변량 시계열 데이터를 통해 이상 상태를 확인하고 원인을 파악하여 알맞은 조치를 취함으로써 생산 능력과 이윤을 극대화한다.

공정의 이상 상태 정상화 또는 공정 효율 상승을 위해서는 공정의 상태를 예측하고 원인을 분석하는 과정이 필요하다. 공정의 상태를 파악하기 위하여 서포트 벡터 머신(Support Vector Machine), 의사 결정 나무(Decision Tree) 등의 기계학습(Machine Learning) 모델과 함께 합성곱 신경망(Convolutional Neural Network), 순환 신경망(Recurrent Neural Network) 등의 인공 신경망(Artificial Neural Network) 모델을 통한 분류와 예측 방법 연구가 활발히 진행되고 있다.

하지만 공정의 이상, 효율의 하락 등을 예측하고 알맞은 조치를 통해 정상화 및 효율 상승으로 이루어지기 위해서는 예측과 분류의 결과가 해석 가능(Interpretable)해야 한다. 또한 지도학습(Supervised Learning)을 위한 초기 데이터의 분류 작업과 새로운 이상 상태에 대한 해석은 공정 엔지니어들의 직접적인 데이터 관찰이 필요하다.

본 연구에서는 공정 데이터의 합성곱 신경망을 통한 학습 성능과 데이터 분석가의 해석 능력을 향상시키기 위하여 다변량 시계열 데이터의 변수 인덱싱을 이용한 전처리 방법을 제안하였다. 0-1 스케일링한 여러 가지 시계열 데이터의 상관 계수를 확인하고 이웃한 데이터 간 상관 계수가 최대화되도록 재배열하고 반전시키는 변수 인덱싱을 하였다.

다변량 시계열 데이터를 흑백의 단 채널 형태로 그림 1.1과 같이 시각화하였을 때 인덱싱 전의 경우 시계열 데이터 간 상관 계수가 낮아 시계열 간 상호 작용을 확인하기 어렵다. 하지만 인덱싱 후의 경우 비슷한 값을 가지는 데이터가 이웃에 배치되어 상호 작용과 그 변화를 쉽게 확인할 수 있다.

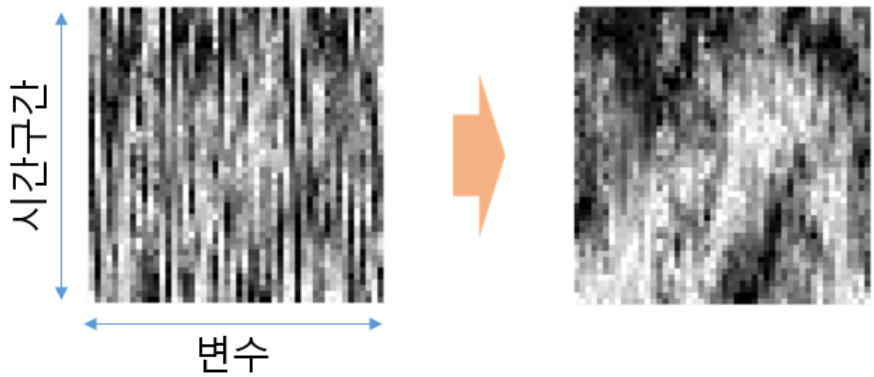

Figure 1.1: 단 채널(흑백)로 시각화한 다변량 시계열 데이터(인덱싱 전(좌)과 후(우))

인덱싱 효과를 검증하기 위해 제철 공정에서 발생하는 다변량 시계열 데이터를 이용한 합성곱 신경망 예측 실험을 하였다. 공정의 효율을 예측하는 실험과 임의로 추가한 노이즈를 감지하는 실험을 인덱싱 전, 후 결과에 대해 비교하였다. 그리고 추가한 노이즈의 위치를 데이터 분석가가 인지할 수 있는지에 대한 실험으로 추가 비교하였다.

본 논문의 구성은 다음과 같다. 2장에서는 다변량 시계열 데이터의 합성곱 신경망 적용 방법에 대한 관련 연구를 살펴보고 3장에서는 제안한 재배열과 반전을 이용하는 변수 인덱싱 방법에 대하여 자세히 설명한다. 4장에서는 인덱싱 전, 후 데이터에 대한 합성곱 신경망과 데이터 분석가 인지실험에 대하여 평가하고 마지막으로 5장에서 본 연구의 의의에 대하여 논한다.

제 2 장 선행연구

본 연구는 다변량 시계열의 전처리를 통한 결과에 대해 합성곱 신경망과 인간의 인지 성능 실험을 통해 성능을 비교하였다. 다변량 시계열 데이터를 합성곱 신경망으로 분류하여 예측한 사례를 소개한다.

2.1 다채널 단변량 시계열 합성곱 신경망(Zheng et al. (2014))

Zheng et al.(2014)[5]은 그림 2.1과 같이 다변량 시계열 데이터의 각 변수를 단 변량 시계열 데이터 형태로 합성곱 신경망을 적용하였다. 각 시계열 데이터를 별도의 채널로 구성한 합성곱 신경망을 통해 추출한 결과를 모두 연결한 다층 퍼셉트론 구조를 통하여 분류를 진행하였다.

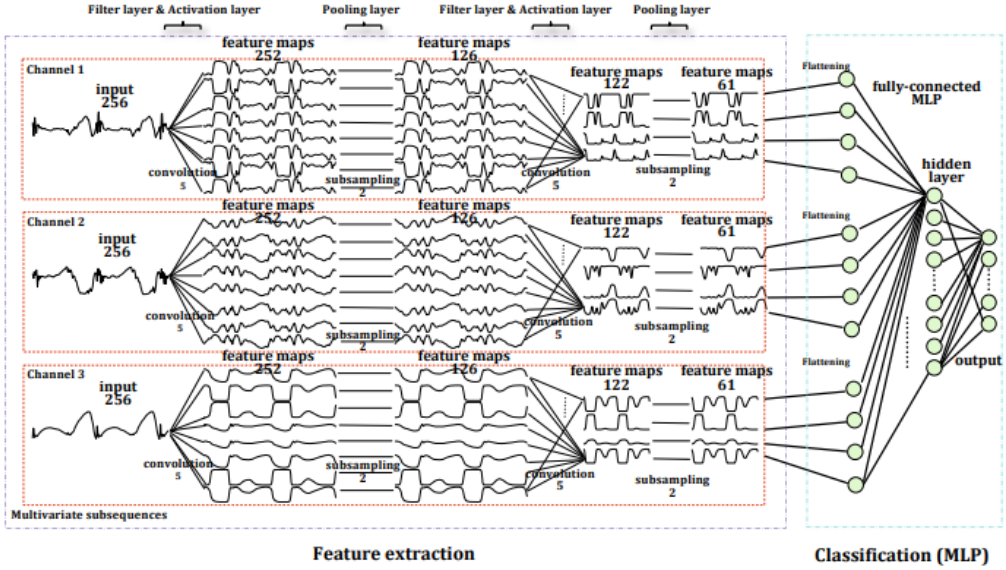

Figure 2.1: 다채널 단변량 시계열 합성곱 신경망 구조(Zheng et al.(2014))

채널마다 별도의 합성곱 신경망을 적용하여 각 시계열 데이터마다의 특징을 추출하고 이를 종합하여 분류하는 데 사용한다.

시계열 데이터마다 특징을 추출하고 이를 종합하기 때문에 다른 시계열 간의 상호관계에 의한 특징을 직접적으로 추출하지 않는다. 또한 데이터 분석가가 분류되는 원인을 찾거나 새로운 유형을 나누기 위해서 각각의 시계열 데이터 그래프를 통해 찾아야 한다.

2.2 Fault Detection and Classification 합성곱 신경망(Lee et al. (2014))

Lee et al.(2017)[1]은 공정에서 발생하는 다변량 시계열 데이터를 그림 2.2와 같은 구조의 FDC(Fault Detection and Classification) 합성곱 신경망을 이용하여 분류하였다.

합성곱 신경망에서 사용하는 수용영역(Receptive field)을 정사각 형태가 아닌 시계열 데이터(센서 데이터)의 숫자가 폭이 되도록 사용하였다. 수용영역을 시간 방향으로만 이동해가면서 값을 추출한다.

$$y_{ij} = \sigma(\sum_{r=1}^{F}\sum_{c=1}^{F} w_{rc}x_{(r+i\times s_r)(c+j\times S)} + b), \quad 0 \le i \le \frac{H-F}{S}, 0 \le j \le \frac{K-F}{S} \tag{2.1}$$

$$y_{i} = \sigma(\sum_{r=1}^{F_r}\sum_{c=1}^{W} w_{rc}x_{(r+i\times s_r)(c)} + b), \quad 0 \le i \le \frac{H-F_r}{S_r} \tag{2.2}$$

일반적인 합성곱 필터의 정사각 수용영역의 가로, 세로 크기 F, 입력층의 가로, 세로 차원 수 W, H, 수용영역의 보폭 S, 가중치 w, 바이어스 b, 활성 함수 σ에 대한

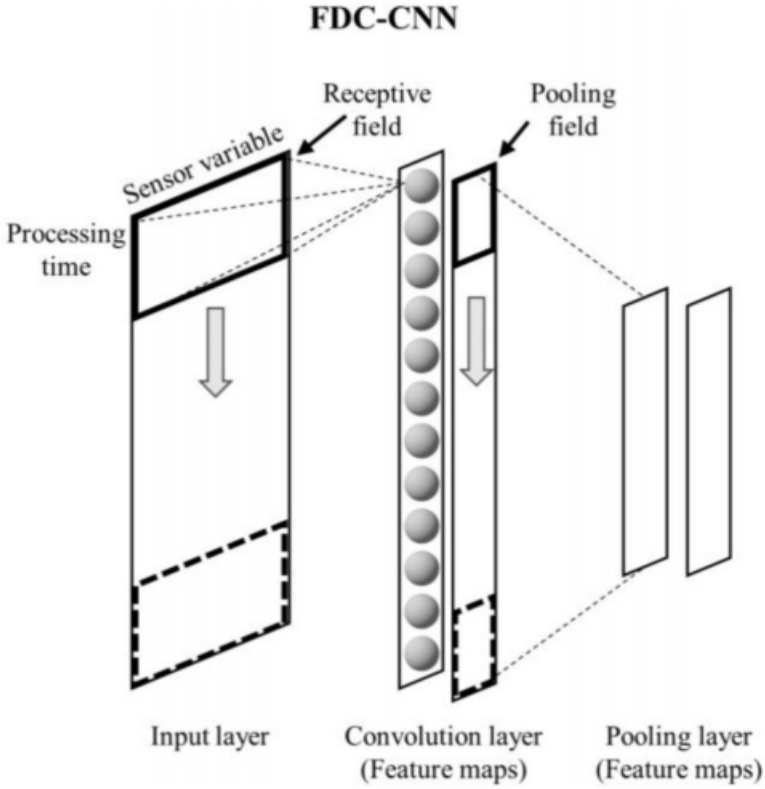

Figure 2.2: FDC-합성곱 신경망 구조(Lee et al.(2017))

출력값은 다음 식과 같다.(2.1) Lee et al.(2017)의 합성곱 연산은 수용영역의 크기가 정사각 형태가 아니고 시간 구간(Time Step) 방향으로만 스캔을 하기 때문에 수용영역의 세로(시간 구간) 방향 크기 F_r, 시간 구간 방향 보폭 S_r에 대한 출력값은 다음 식과 같다.(2.2) 활성 함수 σ는 ReLU 함수를 사용하였다.

FDC 합성곱 신경망과 일반적인 합성곱 신경망 모델을 테스트한 결과 FDC 합성곱 신경망의 분류 성능이 더 좋다. 수용 영역이 모든 센서에 대하여 시간 구간별로 하나의 추출 값을 내보낸다. 분류의 원인이 되는 시간 구간 탐색은 용이하나 분류의 원인이 되는 센서를 추적하기는 어렵다.

2.1 절과 2.2 절의 데이터 분류 방법을 데이터 분석가가 시계열 데이터 간 상호 관

계에 의한 새로운 유형을 분류하려면 각각의 데이터 그래프를 비교하여야 한다. 또한 기계 학습에 의한 원인 시간 구간과 원인 센서 추적을 동시에 하기 어렵다.

제 3 장 제안하는 모델

무작위로 상승, 증가하도록 생성한 다변량 시계열 데이터를 0-1 스케일링 하면 그림 3.1과 같게 된다. 이웃한 행, 열 간 상관 계수를 계산하면, 시간의 전후 관계를 나타내는 행 간의 상관 계수는 0.89로 높지만 서로 다른 센서의 값을 나타내는 열 간의 상관 계수는 0.06으로 낮다. 즉, 다변량 시계열 데이터의 시간 방향(열 방향)으로 이웃한 값은 비슷한 크기를 갖게 되지만, 시계열 데이터 간(서로 다른 센서 값, 열 간)을 나타내는 행 방향으로는 비슷한 크기를 갖지 않는다.

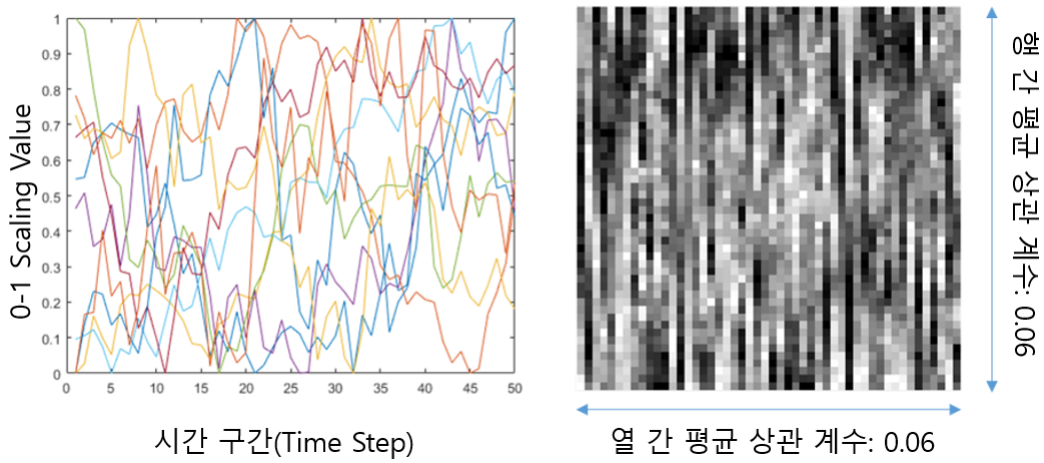

Figure 3.1: 0-1 스케일링한 다변량 시계열 데이터의 그래프(좌)와 시각화(우)

이미지의 경우 이웃한 위치에 관계 없이 픽셀 간 상관 계수가 높게 나타난다. 물체를 식별하기 유리한 높은 해상도 일수록 높은 상관 계수를 가진다. 그림 3.2는 이미지의 해상도에 따른 Red, Green, Blue 채널 별 이웃한 픽셀 간의 상관 계수를 나타낸다.

다변량 시계열 데이터의 시간 방향의 이웃한 데이터 간 상관 계수는 이미지와 비슷하지만 이웃한 시계열 간의 상관관계는 거의 없다. 각 시계열 데이터의 정보를 잃지

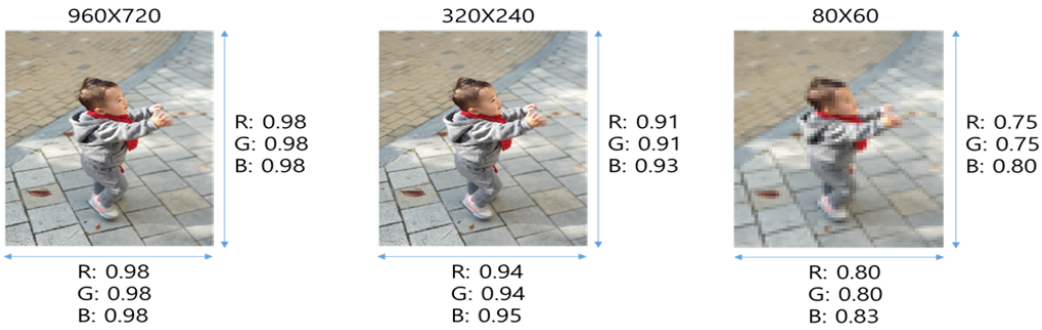

Figure 3.2: 해상도에 따른 이미지 채널 별 상관계수

않도록 새롭게 인덱싱하여 시계열 데이터 간 상관 계수를 높였다.

3.1 데이터의 재배열 및 반전을 통한 인덱싱

다변량 시계열 데이터의 시계열 데이터 간 상관관계를 높이기 위해 두 가지 인덱싱 방법을 적용하였다. 시계열 데이터(열)를 상관 계수가 높아지도록 그림 3.3과 같이 비슷한 값을 가지는 열끼리 이웃하도록 재배열하고, 이웃 간 상관 계수의 부호를 바꾸도록 그림 3.4와 같이 0-1 스케일링한 값 x를 $1 - x$로 변환하는 반전을 하였다.

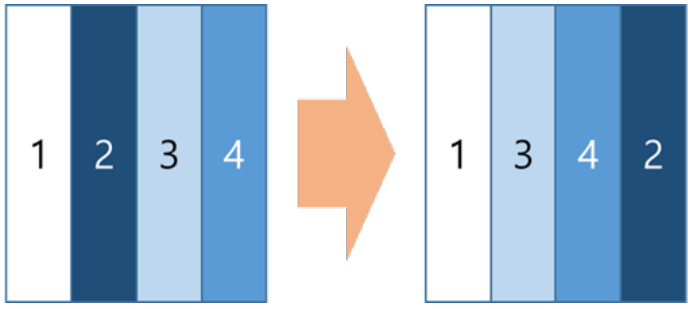

Figure 3.3: 다변량 시계열 데이터의 재배열

n개 열에 대한 최적 재배열을 활용한 인덱싱의 경우, 이웃한 열 간 상관 계수 합은 상관 계수 행렬(Correlation Matrix)에서 행과 열이 중복되지 않는 $n - 1$개의 조합 중 최대의 합으로 계산된다. 하지만 최적의 재배열과 반전을 동시에 활용한 인덱싱의 경우,

Figure 3.4: 시계열 데이터의 반전

상관 계수 행렬 원소의 절댓값 조합 중 합이 최대가 된다.

최적 인덱싱을 구하기 위해 완전 그래프의 경로 탐색 문제로 변환하여 정수최적화를 활용하였다. 3.1.1 절에서는 완전그래프로의 변환을 소개하고, 3.1.2 절에서 정수최적화 해법을 소개한다.

3.1.1 완전 그래프 문제로의 전환

다변량 시계열 데이터의 인덱싱 문제를 그림 3.5와 같은 완전 그래프로 변환하였다. 완전 그래프의 노드는 각각의 시계열 데이터를 나타내고, 각 링크는 상관계수를 표현한다. 이때 노드 i 와 j 간 링크 비용 c_{ij} 는 연결된 노드 간의 상관 계수 r_{ij}의 절댓값에 -1 을 곱한 $-|r_{ij}|$ 가 된다.

다변량 시계열 데이터의 인덱싱 문제는 완전 그래프의 해밀턴 경로 문제(모든 노드를 지나는 최단 경로 문제)로 바꿀 수 있다. 각 시계열 데이터의 배열 순서는 출발 노드부터 차례대로 왼쪽부터 오른쪽으로 배열하게 된다. 출발 노드에 해당하는 시계열

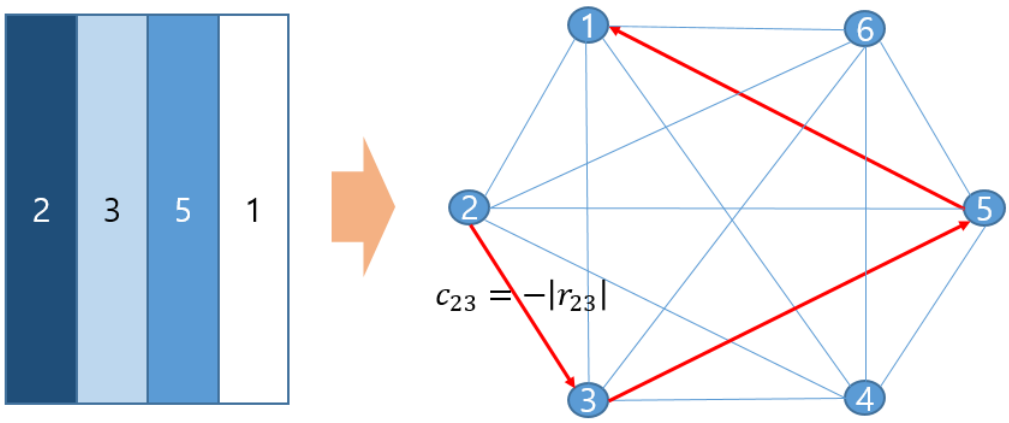

Figure 3.5: 완전 그래프 문제로의 변환

데이터가 가장 왼쪽에 위치하고 도착 노드에 해당하는 시계열 데이터가 가장 오른쪽에 위치하게 된다. 그림 3.5와 같이 2, 3, 5, 1 순서의 시계열 데이터의 배열 순서는 2, 3, 5, 1 순서의 경로로 나타낼 수 있다.

이때, 인접한 시계열 데이터의 상관 계수 합이 최대가 되는 배열은 최소 비용 경로 탐색 문제가 된다. 전체 노드를 지나는 경로의 총비용은 $(-1) \times$(이웃 열 간 상관 계수의 합)이 된다.

3.1.2 정수최적화

다변량 시계열 데이터의 인덱싱 문제를 그래프 문제 변환한 뒤, 최적해를 구하기 위하여 정수최적화 문제로 수식화 하였다.[2][3]

$$\underset{x,u}{\text{minimize}} \quad \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} c_{ij} x_{ij} \tag{3.1}$$

$$\text{subject to} \quad x_{ij} \in \{0, 1\} \tag{3.2}$$

$$\sum_{i=1,i\neq j}^{n} x_{ij} \leq 1 \qquad j = 1, 2, ..., n \tag{3.3}$$

$$\sum_{j=1,j\neq i}^{n} x_{ij} \leq 1 \qquad i = 1, 2, ..., n \tag{3.4}$$

$$\sum_{i=1}^{n} \sum_{j\neq i,j=1}^{n} x_{ij} = n - 1 \tag{3.5}$$

$$\begin{gathered} u_i - u_j + n x_{ij} \leq n - 1 \qquad i \neq j \\ 0 \leq u_i \leq n - 1 \end{gathered} \tag{3.6}$$

시계열 데이터 n개에 대해 인덱싱하기 위한 최소 비용 경로의 목적함수는 (3.1)과 같다. c_{ij}는 i 노드와 j 노드 간의 상관 계수 r_{ij}의 절댓값에 -1을 곱한 값이 되고, x_{ij}는 i 노드와 j 노드를 연결하는 링크의 선택 여부를 나타낸다. 제약식 (3.2)에서 1은 해당 링크를 선택한 경로를, 0은 해당 링크를 선택하지 않은 경로를 나타낸다.

제약식 (3.3), (3.4)에 의해 각 노드는 링크의 출발점으로 0회 또는 1회, 도착점으로 0회 또는 1회 선택될 수 있다. 제약식 (3.5)는 경로에 해당하는 링크가 $n-1$개 선택되는 것을 의미한다. 가장 왼쪽 열에 해당하는 출발 노드와 가장 오른쪽 열에 해당하는 도착 노드를 제외하면 모든 노드는 링크의 출발점과 도착점으로 각 1회 씩 선택된다. 출발 노드의 경우 링크의 출발점으로 1회 선택되지만 도착점으로 선택되지 않고, 도착 노드의 경우 링크의 도착점으로 1회 선택되지만 출발점으로 선택되지 않는다. 제약식 (3.3)은 출발 노드에 해당하는 j의 경우만 0이 되고 나머지는 1이 되며, 제약식 (3.4)는 도착 노드에 해당하는 i의 경우만 0이 되고 나머지는 1이 된다.

제약식 (3.6)은 회로의 생성을 막는다. 차수가 0 또는 1인 부분 그래프는 경로가

회로를 요소로 가져야 하는데 식 (3.6)에 의해 회로를 가질 수 없으며, 식 (3.5)에 의해 정확히 n-1개의 호를 가지기 때문에 모든 마디를 지나는 회로가 되어야 한다. u_i는 i 번째 시계열 데이터의 배열 위치를 뜻한다. i 노드와 j 노드 간의 링크가 선택되지 않은 경우 $u_i - u_j$의 최댓값은 $n-1$로 제약식은 성립하고, 링크가 선택되는 경우 $x_{ij} = 1$ 이므로 $u_i < u_j$가 되어야 한다. 이를 만족하기 위한 u_i는 가장 왼쪽에 배치되는 열부터 0, 1, ..., $n-1$이 된다.

3.1.3 최적해 적용

3.1.2 절 정수최적화 문제의 최적해 u_i^*로부터 재배열 순서가 결정된다. 각 열의 반전여부를 결정하기 위하여 $u_i^* = k$가 되는 값을 이용하여 $o_k = i$를 도입한다. i번 째 열의 반전여부 s_i는 식 (3.7)과 같다. 이때 $s_i = 1$인 경우 반전을 하지 않고 $s_i = -1$인 경우 해당 열을 반전시킨다.

$$s(i) = \begin{cases} 1 & for \quad i = 1, \\ \prod_{k=1}^{i-1}\{\mathbb{1}(r_{o_{k-1},o_k} > 0) - (1 - \mathbb{1}(r_{o_{k-1},o_k} > 0))\} & for \quad i = 2, 3, ..., n. \end{cases} \tag{3.7}$$

$s(1) = 1$로 첫 번째 열은 반전하지 않고 각 열은 왼쪽 열의 반전 여부와 상관 계수로 반전을 결정한다. 왼쪽 열과의 상관계수가 양수인 경우 $s(i) = s(i-1)$이 되고 양수가 아닌 경우 $s(i) = -s(i-1)$이 된다.

아래의 X_i는 인덱싱 전 i번째 시계열 데이터를 나타내고, X'_i는 이웃한 열 간 상관계수의 합이 최대가 되도록 인덱싱한 시계열 데이터의 i번째 열을 나타낸다.

$$X_i^{'} = \mathbb{1}(s(i) = 1)X_{o_i} + (1 - \mathbb{1}(s(i) = 1))(-X_{o_i} + 1) \tag{3.8}$$

재배열과 반전 여부를 적용하여 0-1 스케일링한 다변량 시계열 데이터 X를 인덱싱한 X'은 식 (3.8)가 된다.

제 4 장 실험 및 결과

인덱싱 효과를 검증하기 위하여 3가지 실험을 진행하였다. 첫째로 이웃 간 상관 계수의 합의 변화를 확인하였고, 둘째로 합성곱 신경망을 통한 분류 성능 차이를 비교하였다. 마지막으로 데이터의 노이즈를 인간이 인지하는데 개선이 있는지 실험을 통하여 확인하였다.

실험에 사용한 데이터는 제철 공정에서 측정되는 센서 데이터를 이용하였다. 인지 실험에는 데이터 분석 전문가 35명에 대하여 실험을 진행하였다. 실험에 사용된 정수 최적화 문제를 풀기 위하여 MATLAB의 혼합 정수 선형 계획법 솔버[4]를 활용하였다.

4.1 상관 계수 변화 실험

상관 계수 변화 실험은 두 가지로 진행하였다. 랜덤으로 상승 하락을 반복하는 데이터에 대한 상관 계수 변화를 확인하고 제철 공정에서 발생하는 센서 데이터에 대한 상관 계수 변화를 확인하였다.

4.1.1 랜덤 생성 데이터 변화

랜덤하게 증가하고 감소하는 데이터를 두 가지 생성하여 인덱싱 전, 후 결과를 비교하였다. 50개의 시간 구간을 가지는 50개의 시계열 데이터와 100개의 시간 구간을 가지는 100개의 시계열 데이터를 인덱싱하였다.

인덱싱 결과는 그림 4.1과 같다. 이웃 간 상관 계수의 평균값이 50×50의 경우 0.73으로 100×100의 경우 0.76으로 크게 증가하는 것을 확인하였다.

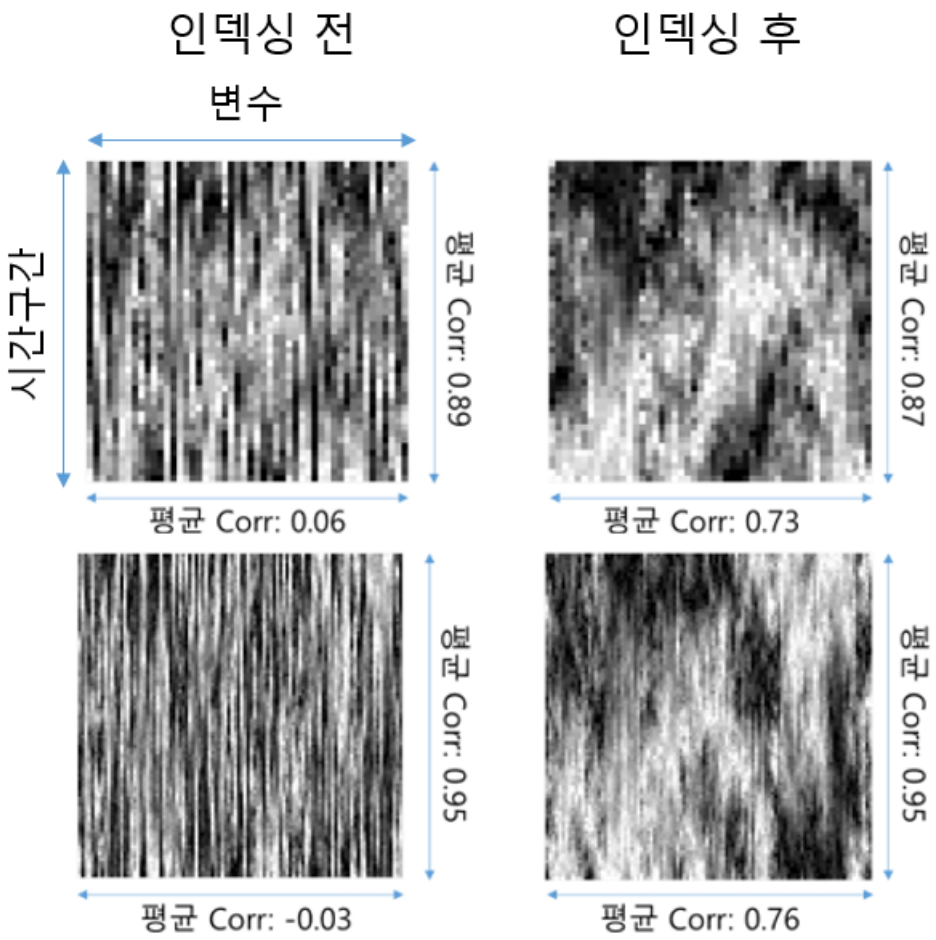

Figure 4.1: 랜덤 생성 데이터 상관계수 변화(상: 50×50, 하: 100×100)

4.1.2 제철 공정 데이터 변화

제조 공정에서 발생하는 센서 데이터를 트레이닝, 테스트 데이터로 나누어 인덱싱하교 결과를 비교하였다. 총 35개의 센서 데이터에 대하여 앞의 28,600시간 구간의 트레이닝 데이터와 뒤의 28,580시간 구간의 테스트 데이터로 구분하였다. 트레이닝 데이터의 인덱싱을 테스트 데이터에도 적용하여 결과를 비교하였다.

인덱싱 전, 후의 결과를 이미지로 표현하면 그림 4.2와 같게 된다. 트레이닝 데이터의 인덱싱 전, 후 이미지 형태와 트레이닝 데이터의 인덱싱을 적용한 테스트 데이터의

Figure 4.2: 제철 공정 데이터 인덱싱 결과(트레이닝(상)과 테스트(하))

이미지의 형태가 비슷하게 바뀐 모습이다. 인덱싱 전, 후의 상관 계수는 표 4.1과 같다. 트레이닝 데이터의 상관 계수가 0.54로 증가하였고, 테스트 데이터도 비슷한 수준인 0.53으로 증가하였다.

Table 4.1: 제철 공정 데이터 인덱싱 결과(열 간 상관 계수)

	인덱싱 전	인덱싱 후
트레이닝	0.05	0.54
테스트	0.05	0.53

제조 공정의 데이터의 경우, 센서 값 사이의 상관 계수가 동일한 공정 조건에서 일정하게 유지되는 것을 확인할 수 있다. 트레이닝 데이터의 인덱싱 효과가 테스트 데이터에도 유효함을 확인하였다.

4.1.3 금융 데이터 변화

금융에 사용되는 다변량 시계열 데이터 중 하나인 세계 주가지수 데이터에 대하여 인덱싱을 진행하였다. 인덱싱 결과는 그림 4.3과 같다. 트레이닝 데이터를 인덱싱 하였을 때 이웃 열 간 상관 계수가 0.52에서 0.86으로 크게 증가하였으나, 이 인덱싱을 테스트 데이터에 적용하였을 때 이웃 열 간 상관 계수는 0.43에서 0.57로 소폭 증가하였다.

금융 데이터의 경우 시간이 지나게 되면 시계열 데이터 간 상관 계수가 변하게 된다. 트레이닝 데이터의 인덱싱을 통한 이웃 열 간 상관 계수의 증가 효과가 테스트 데이터에 동일하게 적용되지 않는다.

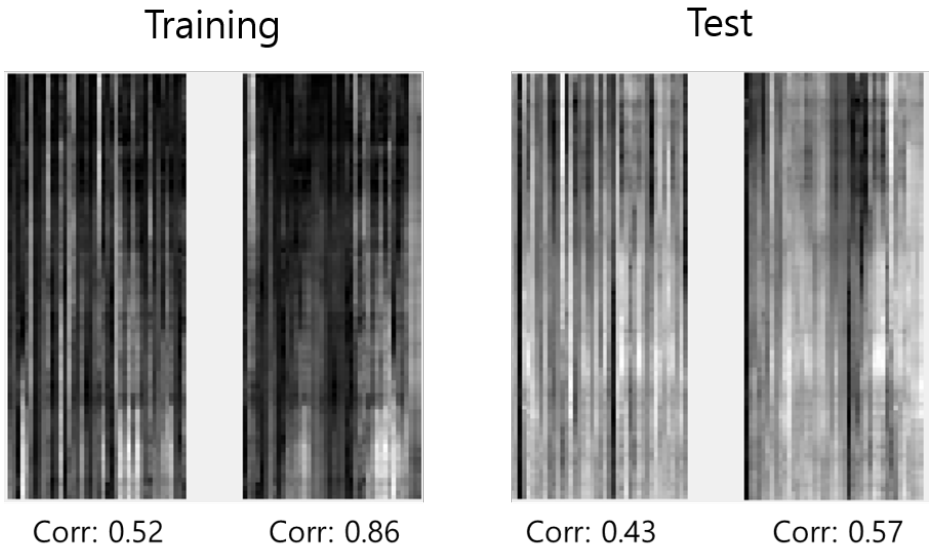

Figure 4.3: 세계 주가지수 데이터 인덱싱 결과(트레이닝(좌)과 테스트(우))

4.2 합성곱 신경망 적용 실험

인덱싱 전, 후 데이터에 대하여 합성곱 신경망을 이용한 2가지 분류 모델을 이용하여 비교하였다. 첫째로 4.1.2 절에서 인덱싱한 제철 공정 데이터를 이용하여 공정의 효율을 예측하는 분류 하는 실험을 진행하였고, 둘째로 제철 공정 데이터 중 절반에 노이즈를 추가하고 노이즈 여부를 분류하는 실험을 진행하였다.

두 가지 실험 모두 35개의 0-1 스케일링한 시계열 데이터 5분 간격의 20 시간 구간을 사용하였다. 700픽셀(20×35)의 1채널(흑백) 이미지를 합성공 신경망에 적용하였다. 사용한 합성곱 신경망 구조는 그림 4.4과 같다. 3개의 3×3 합성곱 층을 사용하고, 한 층의 Fully Connected 층을 거쳐 Softmax를 통해 출력하도록 하였다. 학습 시 0.8의 확률로 링크를 유지하도록 각 층에 Dropout을 적용하였다. 각 층의 활성함수로는 ReLU 함수를 사용하였다. Batch의 크기는 128을 사용하였고 Root Mean Square Propagation 알고리즘으로 학습하였다.

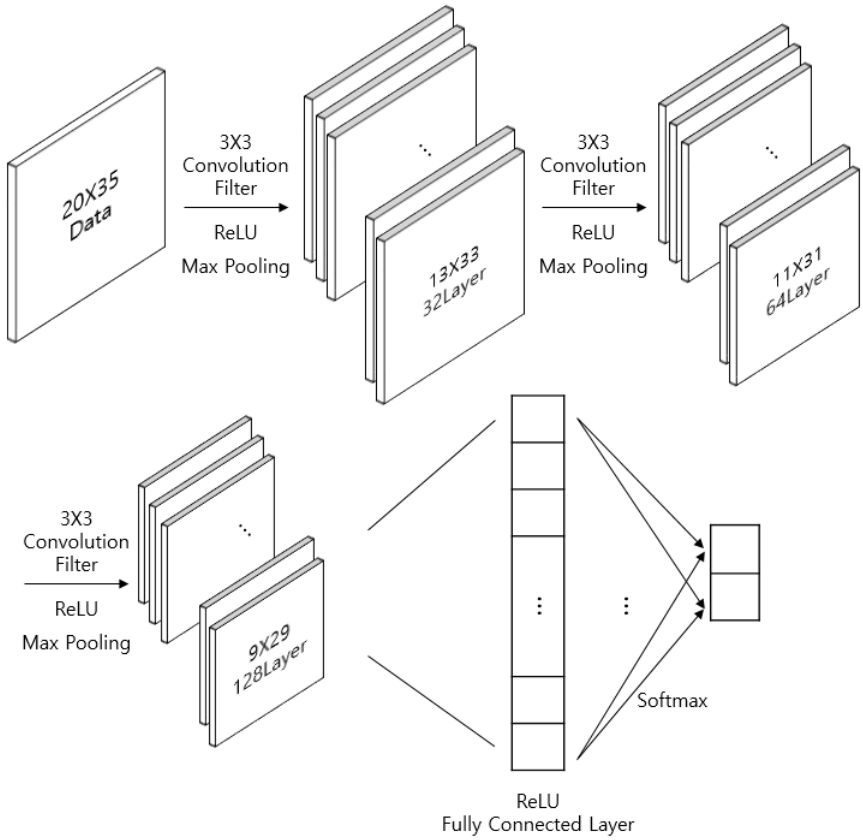

Figure 4.4: 사용한 합성곱 신경만 구조

4.2.1 공정 데이터 효율 예측

제철 공정 센서에서 발생하는 다변량 시계열 데이터를 통하여 공정의 생산 효율을 예측하는 분류 실험을 진행하였다. 단위 당 생산 가격을 기준으로 상위 50%와 하위 50%를 분류하고 이를 트레이닝과 테스트 세트로 구분하여 학습하고 평가하였다.

100 Epoch 학습을 통해 트레이닝한 결과 트레이닝 데이터와 테스트 데이터의 적용 결과는 표 4.2와 같다. 테스트 데이터에 대한 인덱싱 전, 후 예측 정확도는 52.0%에서 57.1%로 향상 되었다.

다변량 시계열 데이터의 인덱싱이 합성곱 신경망의 성능을 향상 시켰다. 공정 데이

터의 경우 상관 계수가 높은 센서 값들로 부터 정보를 추출하는 것이 분류 성능 향상에 도움이 된다.

Table 4.2: 공정 데이터 효율 예측 결과 (정확도(%))

	인덱싱 전	인덱싱 후
트레이닝	55.2	60.1
테스트	52.0	57.1

4.2.2 공정 데이터 노이즈 분류

앞에서 사용한 제철 공정 데이터의 일부에 노이즈를 추가하고 이 노이즈의 유무를 분류하는 테스트를 진행하였다. 4.1.2 절에서 인덱싱 한 20×35 사이즈의 700픽셀 이미지 데이터 57199개 중 절반인 28600개에 노이즈를 추가하였다.

0-1 스케일링된 700개의 픽셀 중 열 방향으로 연달아 세 픽셀의 값을 0.15값 만큼 더하거나 빼는 방법으로 노이즈를 추가하였다. 추가한 형태는 그림 4.5와 같다. 원본(왼쪽) 데이터에 3×1 크기의 노이즈를 추가(가운데)하고, 위치를 표시(오른쪽) 하였다.

노이즈 추가 유무를 찾는 그림 4.4와 같은 구조의 합성곱 신경망 구조를 이용하여 분류 하였다. 40000개의 트레이닝 데이터를 이용해 학습하고, 17199개의 테스트 데이터를 이용해 평가하였다.

Table 4.3: 노이즈 유무 예측 결과 (정확도(%))

	인덱싱 전	인덱싱 후
트레이닝	91.2	93.0
테스트	84.2	85.1

총 200 Epoch의 학습을 통한 결과는 표 4.3과 같다. 인덱싱 후의 노이즈 유무에 대한 예측 정확도가 상승하는 것을 확인할 수 있다. 테스트 데이터의 경우 84.2%에서 85.1%로 소폭의 개선을 보였다. Epoch에 따른 분류 정확도와 손실함수 값은 그림 4.6

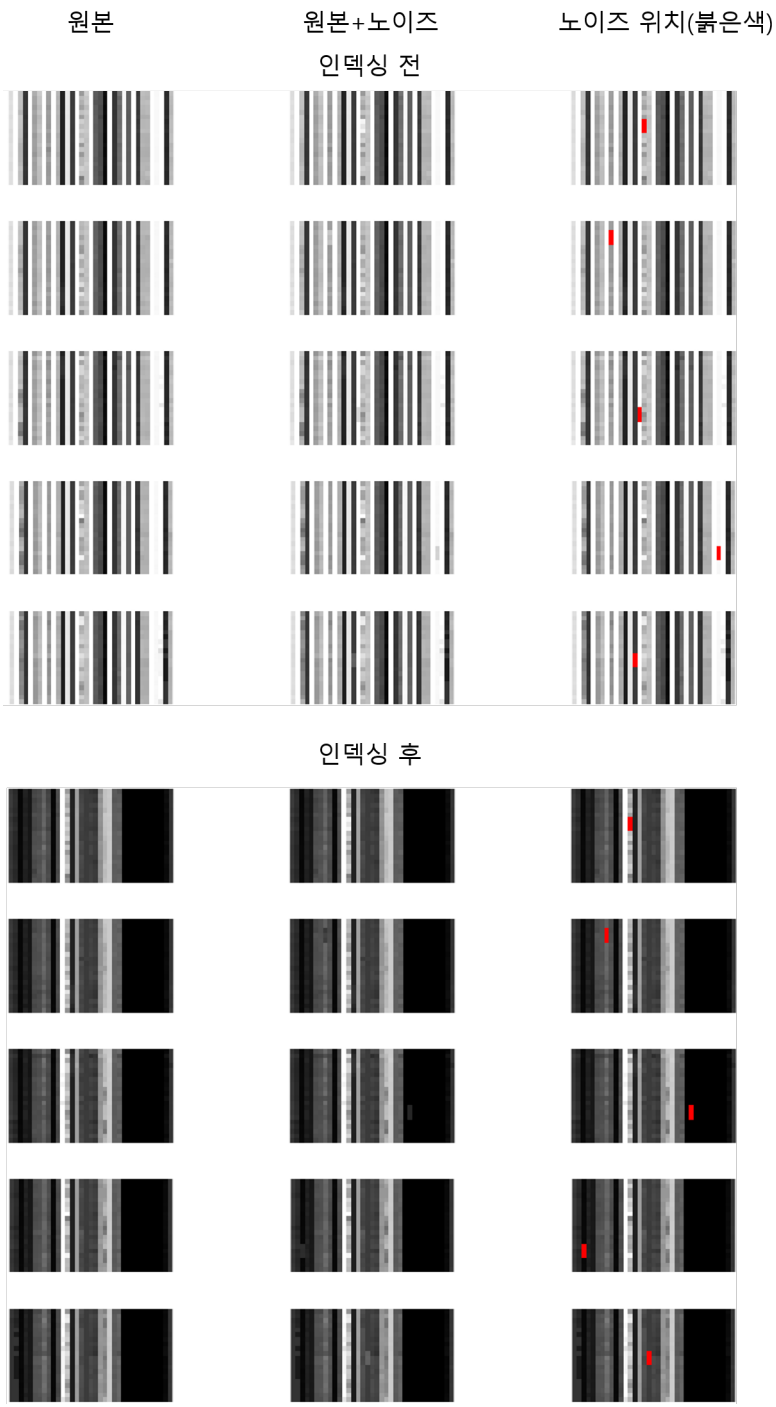

Figure 4.5: 인덱싱 전(상), 후(하) 데이터에 추가한 노이즈

과 같다. 정확도와 손실함수 값이 인덱싱 전은 80 Epoch 근처에서 개선되기 시작하는 것을 확인할 수 있지만, 인덱싱 후의 경우 10 Epoch 미만에서 빠르게 개선되기 시작하는 것을 확인할수 있다. 다변량 시계열 데이터의 인덱싱이 합성곱 신경망의 학습속도를 향상 시키는 것을 확인하였다.

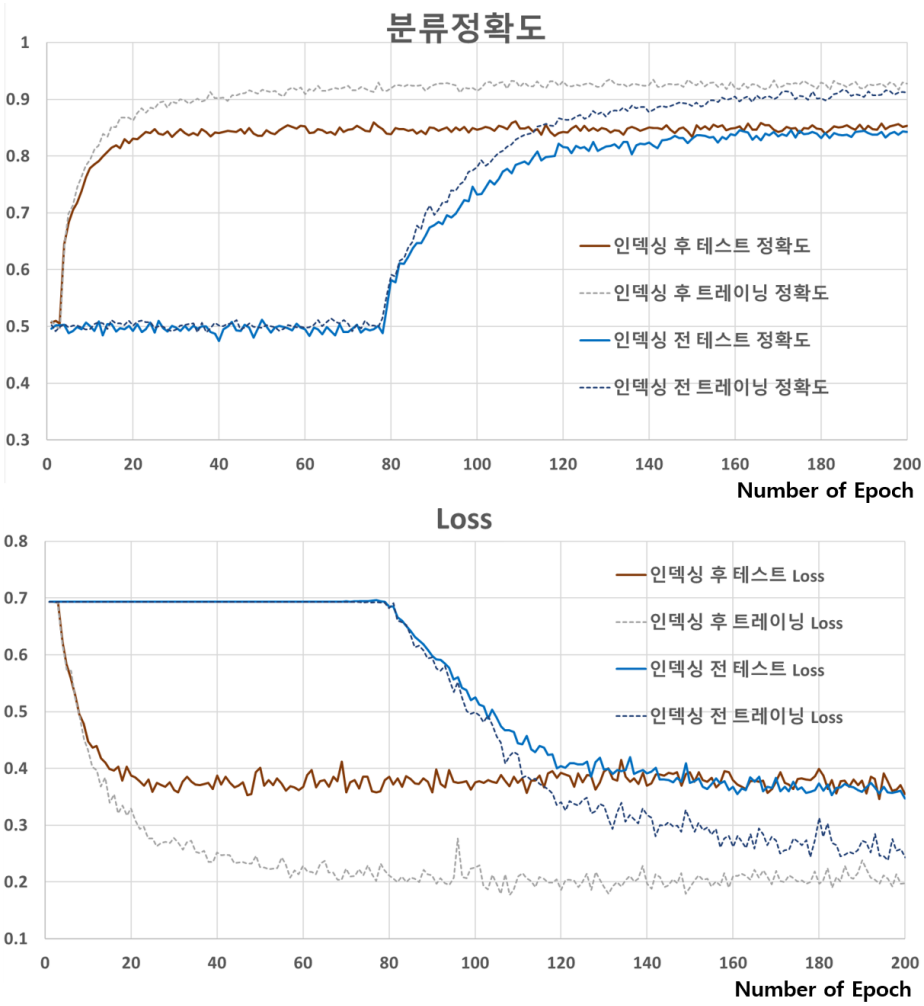

Figure 4.6: 노이즈 유무 예측 Epoch에 따른 분류 정확도(상)와 손실함수 값(하)

4.3 노이즈 인지 실험

4.2.2 절에서 추가한 노이즈를 인간이 찾아내는 실험을 진행하였다. 무작위로 선택한 노이즈가 추가된 10개의 데이터의 인덱싱 전과 후 이미지 총 20를 주고 3×1 크기의 노이즈 위치를 찾아 표시하도록 하였다. 총 35명의 데이터 분석 업무 또는 연구를 하는 사람을 대상으로 실험을 진행하였다.

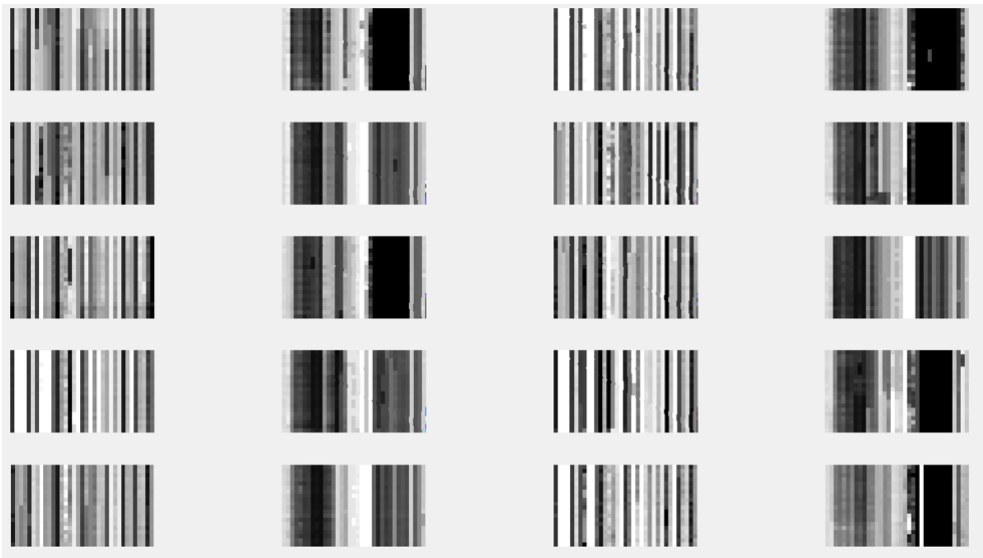

Figure 4.7: 인지실험 테스트 예시

테스트는 그림 4.7과 같이 홀수 열에는 인덱싱 전의 이미지, 짝수 열에는 인덱싱 후의 이미지를 순서를 섞어 배열한 뒤, 20개의 이미지에서 노이즈로 생각되는 위치를 찾아 체크하는 형태로 진행하였다.

총 3가지 세트 중 개인당 한가지 세트에 대하여 위치를 표시하도록 하고 그 결과를 확인하였다. 35명이 인덱싱 전, 후에 대하여 정확한 위치를 맞춘 이미지의 수는 표 4.4와 같다. 35명 중, 인덱싱 전보다 인덱싱 후의 정답을 더 맞춘 사람은 총 27명이었고, 인덱싱 전보다 인덱싱 후의 정답을 더 적게 맞춘 사람은 1명이었다. 정답 수의 분포는

Table 4.4: 인지실험 결과(정답 수)

		인덱싱 후 (정답 수(개))									
		1	2	3	4	5	6	7	8	9	10
인덱싱 전 (정답 수(개))	1										
	2							1			
	3			1			1			1	
	4				3	1	5	3	1	3	
	5					1		2		2	
	6					1	1	1	1	1	
	7							1	3	1	
	8										
	9										
	10										

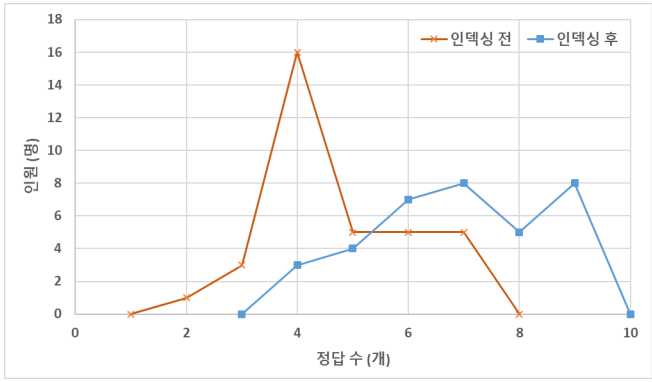

Figure 4.8: 인지실험 결과 분포

그림 4.8와 같고 정답률은 47.1%에서 69.1%로 인덱싱으로 인한 정답률이 22%p 올랐다.

인덱싱을 통하여 데이터 분석가가 데이터의 노이즈 여부를 보다 쉽게 확인할 수 있고 좀 더 많은 노이즈의 위치를 찾아내는 것을 확인하였다. 이웃한 시계열 간 상관계수를 높이는 인덱싱을 통해 인간이 데이터의 변화를 더 쉽게 인지할 수 있다는 것을 알 수 있다.

제 5 장 결론 및 의의

제조 공정에서 이상 발생 시, 이상을 감지하고 원인을 분석하여 그에 알맞은 조치로 대응하게 된다. 이를 위한 초기 이상 여부의 분류와 원인 분석 과정은 공정 엔지니어가 직접 센서를 통해 수집되는 다변량 시계열 데이터를 관찰하여 진행하게 된다. 공정 엔지니어의 데이터 관찰을 돕기 위해 본 연구는 다변량 시계열 데이터가 해석 가능 하도록 전처리하는 방법을 제안하였다.

다변량 시계열 데이터를 시각화하기 위하여 이웃한 시계열 데이터 간의 상관 계수 합이 최대가 되도록 변수를 다음과 같이 인덱싱하였다. 모든 시계열 데이터 간 상관관계를 완전 그래프로 표현하고 모든 노드를 지나는 최단 경로를 탐색하는 문제로 변환하였다. 이를 정수최적화 형태로 수식화하여 최적해를 도출하고 변수들을 새롭게 인덱싱하였다. 그 결과를 이용하여 다변량 시계열 데이터를 이미지 형태로 시각화하였다.

제철 공정 데이터에 대하여 트레이닝과 테스트로 나누어 인덱싱한 결과 트레이닝 데이터의 상관 계수 변화와 테스트 데이터의 상관 계수 변화가 비슷한 것을 확인하였다. 같은 공정 조건으로 공정에서 발생하는 시계열 데이터 간에 상관 계수는 유지되는 것을 확인할 수 있었다. 그리고 인덱싱 전, 후의 결과를 이용하여 합성곱 신경망을 통한 분류와 데이터 분석가들의 인지 실험을 진행하였다. 그 결과 인덱싱한 데이터의 분류 성능과 학습 속도가 향상되는 것을 확인하였고 데이터 분석가가 인덱싱 후 데이터의 변화를 더 정확히 발견하는 것도 확인하였다.

본 연구의 의의는 다음과 같다. 첫 번째로, 공정 데이터 분석을 돕는 전처리 방법을 제안하였다. 센서로부터 생성되는 수십에서 수백 개의 시계열 데이터를 일일이 확인하기는 쉽지 않은 작업이다. 따라서 한눈에 다변량 시계열 데이터를 분석하는 방법이

필요한데 본 연구는 데이터를 시각화하여 하나의 이미지로 표현하고, 그 변화를 더욱 쉽게 인지할 수 있도록 하였다. 그리고 시각화한 이미지를 통해 원인이 되는 위치를 찾고 원인 시간과 원인 센서를 확인할 수 있다. 두 번째로, 합성곱 신경망의 성능을 향상 시켰다. 다변량 시계열 데이터의 경우 합성곱 신경망을 적용할 때 이미지에 주로 사용하는 정사각 수용영역의 합성곱 필터를 사용하지 않는다. 하지만 변수 인덱싱을 통해 전처리한 데이터는 정사각 수용영역을 사용하는 합성곱 신경망의 성능을 향상시켜 이미지에 사용하는 기법들을 적용하기 용이해진다. 세 번째로, 합성곱 신경망 학습 시간을 단축했다. 인덱싱 후의 합성곱 신경망 학습속도가 향상되는 것을 확인하였다. 이는 변수 인덱싱을 통한 전처리가 사전학습(Pre-training)과 같은 효과를 보이는 것이다.

본 연구가 가지는 한계점도 있다. 첫째, 상관관계가 유지되지 않는 다변량 시계열 데이터에 적용이 어렵다. 예를 들어, 금융 데이터의 경우 시간이 지남에 따라 상관관계들이 급격히 변화한다. 이러면 일관된 인덱싱을 유지할 수 없어 매번 새로운 인덱싱을 해야 하고 이는 합성곱 신경망에 적용할 수 없다. 둘째, 하나의 변수가 다른 변수에 일정 시간 구간을 지난 뒤 영향을 주는 경우 이를 이웃하게 배치할 수 없다. 이를 해결하기 위해 변수별로 시간 구간도 일부 조정하는 인덱싱이 필요하다. 위와 같은 문제를 해결할 수 있다면 보다 직관적이고 쉽게 해석 가능한 데이터로의 전처리가 가능할 것이다.

참고 문헌

[1] K. B. Lee, S. Cheon, and C. O. Kim, *A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes*, IEEE Transactions on Semiconductor Manufacturing, 30 (2017), pp. 135–142.

[2] C. E. Miller, A. W. Tucker, and R. A. Zemlin, *Integer programming formulation of traveling salesman problems*, J. ACM, 7 (1960), pp. 326–329.

[3] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1982.

[4] Y. Zhang, *Solving large-scale linear programs by interior-point methods under the matlab environment*, Optimization Methods and Software, 10 (1998), pp. 1–31.

[5] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. Zhao, *Time series classification using multi-channels deep convolutional neural networks*, vol. 8485, Springer Verlag, 2014, pp. 298–310.

Abstract

Feature Indexing for Visualizing Multivariate Time Series Data

Bongjoon Park

Department of Industrial Engineering

The Graduate School

Seoul National University

The manufacturing process produces a variety of time series data from a number of sensors. From this multivariate time series data, the performance and anomalies of the process are predicted. Anomalies of the predicted process lead to appropriate actions through cause analysis. This study proposed a method of preprocessing for visualization by indexing variables so that data analysts can easily analyze causes from multivariate time series data. After each of the multivariate time series data was scaled 0-1, the correlation coefficient between these was converted to a complete graph with the cost of the link. The minimum cost path across all nodes of the complete graph was obtained by integer programming modelling, and variables were indexed by rearranging and reversing based on the optimal solutions. The proposed method was applied to the steel process data to convert it into a single-channel image in a rectangular form and then randomly add noise to find it through a convolutional

neural network and data analyst. For preprocessed data, the classification accuracy and learning speed of the convolutional neural network have been improved, and the data analyst has been able to locate the noise more accurately. The preprocessing method proposed in this study has been identified to improve the performance of the convolutional neural network and the ability of humans to analyze data.

Keywords: Manufacturing Process, Multivariate Time Series, Integer Programming, Convolutional Neural Network, Preprocessing

Student Number: 2017-24488